4

Chapter 4 Review Problems

1. Describe the following plots:

(a)

- curved
- negative
- moderately strong

(b)

- 2 clusters
- positive
- weak

(c)

- curved
- negative, then pos.
- strong

2. Scatterplots show the relationship between 2 quantitative variables

3. Scatterplots can show categorical variables by different colors, shapes of the points.

4. Explanatory variable goes on the X axis, while the response variable goes on the Y axis.

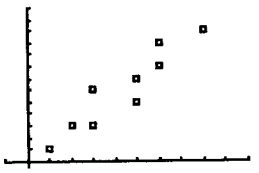
5. What is the range of values that the correlation can take?

$-1 \leq r \leq 1$

6. What type of relationship does r represent?

linear only!

7. For the graph below, what would be the closest approximation to the correlation coefficient?



(a) 0.2
(b) 0.88
(c) –0.9
(d) –0.2
(e) 0
(f) 0.5

8. A plot has a correlation of r = 0.57. I change the units of the x-variable from pounds to kilograms. What happens to the correlation?

nothing. r still equals 0.57

9. What is the coefficient of determination? What is the range of values it can take?

r^2. $0 \le r^2 \le 1$

10 How do we interpret the coefficient of determination?

the ____% of the change in y-variable that is due to/because of the change in the x-variable.

11. Is the correlation affected by outliers?

yes!

12. What is the official name for the line of best fit that we use? (LSR line- what does LSR stand for?)

Least Squares Regression Line

13. The slope of the line of best fit between height in inches (x-variable) and arm span in inches (y-variable) is 1.13. Interpret this slope in context of the problem.

$\frac{1.13 \text{ inches}}{1 \text{ inches}} = \frac{\Delta y}{\Delta x}$

For every 1 inch of arm span, the height increases by 1.13 inches.

14. What is a residual? How do we find/calculate it?

error

actual y value − predicted y value (from LSR line)

15. What is a residual plot?

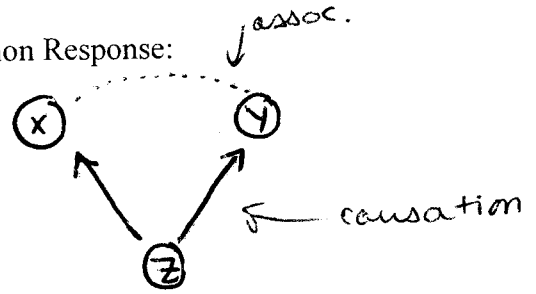
plot of residuals (y) versus x-variable. (x)

16. What does a residual plot tell us? How does it tell us this?

- how well of a fit the linear model is for our data
- If it's scattered, the linear model is good. If there's a pattern, the linear model is not the best model - another would be better.

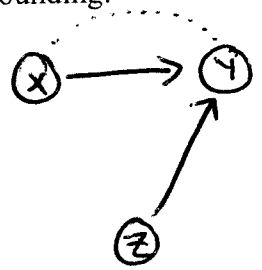
17. Draw diagrams for each of the following and come up with your own example of each

Common Response:



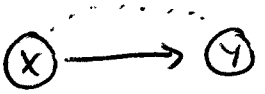
Ex: TV's vs. Life Expectancy.
* the thing causing both to change is $

Confounding:



Ex: Wine vs. heart disease
* wine does lower heart dis.
* other factors affect heart disease (genetics, diet, exercise)

Causation:

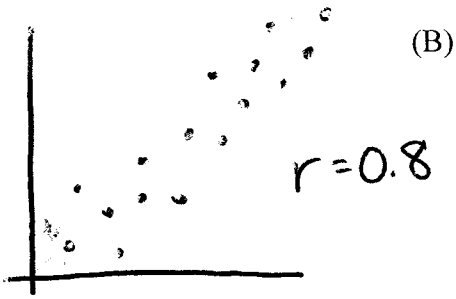


Ex:

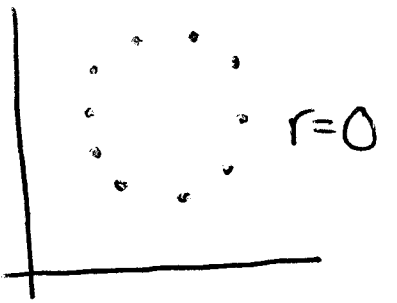
---

18. Match the following pictures to their correlations:

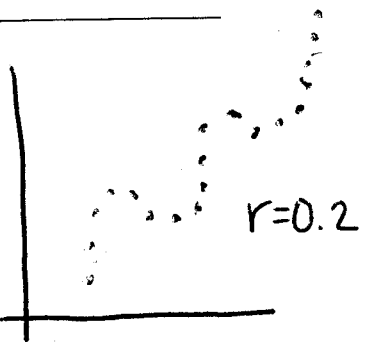
(A)



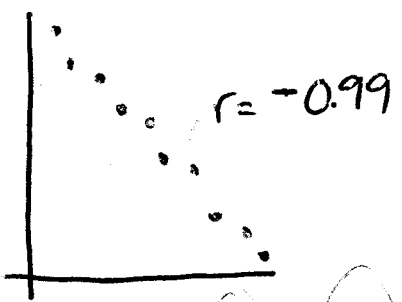
(B)



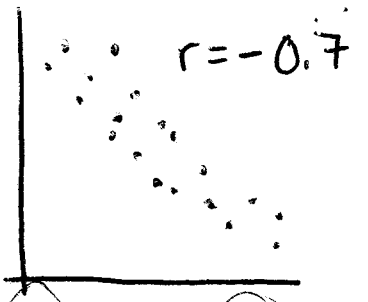
(C)



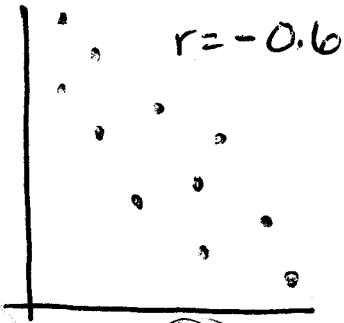
(D)



(E)



(F)



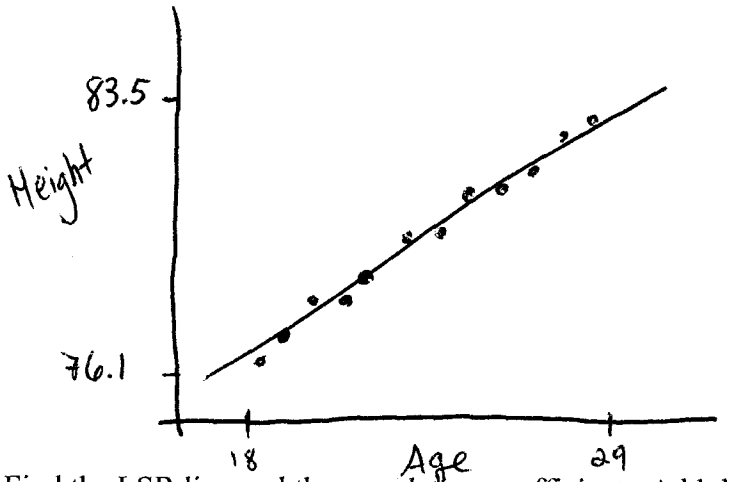
Correlations: **0** B **-0.6** F **0.8** A **-0.99** D **0.2** C **-0.7** E

19. Below is data concerning the mean height of Kalama children. A scientist wanted to look at the effect that age had on the mean height of the children. *For this question, round all numbers to 2 decimal places.*

Age (months)	18	19	20	21	22	23	24	25	26	27	28	29
Height (cm)	76.1	77	78.1	78.2	78.8	79.7	79.9	81.1	81.2	81.8	82.8	83.5

a. Determine the explanatory and response variables Age = expl. Height = resp.

b. Create a scatterplot of the data. Be sure to label the axes. DESCRIBE the plot.



c. Find the LSR line and the correlation coefficient. Add the line to your plot in (b).

$y = 0.63x + 64.93$ $r = 0.994$

d. What percent of the change in the height of Kalama children is explained by the change in their age?

$r^2 = 98.88\%$

e. Interpret the slope of the LSR line in a complete sentence

slope = $\frac{0.63 \text{ cm}}{1 \text{ year}} = \frac{\Delta y}{\Delta x}$ For every 1 year older, the children become 0.63 cm taller.

f. Predict the mean height of a child that is 42 months old (show work!). Is this prediction accurate? Why or why not?

$y = 0.63(42) + 64.93$

* use Y1(42)

$y = 91.60$ cm.

- not accurate.
- 42 months is very far away from original data so we can't trust the line's prediction.

g. Predict the mean height for a child who is 24 months old (show work!)

$y = 0.63(24) + 64.93$ $y = 80.17$ cm

* use Y1(24)

h. Find the error of your predicted value for a child who is 24 months old

$79.9 - 80.17 = -0.27$ cm.

i. Was your prediction an overestimate or an underestimate?

* line predicted higher (over) the actual height.

j. How old is a child expected to be if they are 100cm long? (show work!)

$100 = 0.63x + 64.93$

$x = 55.67$ months

k. Create a residual plot below.

l. What does the plot tell you about your linear model? Explain BRIEFLY.

linear model is a good model because residual plot is scattered.

m. What conclusions can be made from the previous questions? Does the age of a child CAUSE the height of the child? Why or why not?

- It might. High r, r^2, and good residual plot.
- We need more info though. More experiments performed.

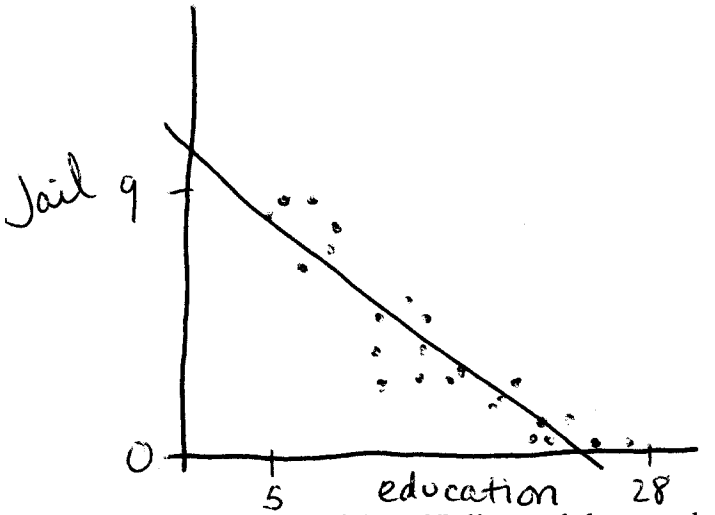
20. Below is data on the years of education versus the years spent in jail by a sample of 20 – 40 year old men. *For this question, round all numbers to 2 decimal places.*

Education (Yrs)	Jail Time (Yrs)
24	0
20	2.1
12	5.2
13	3.6
20	0.5
21	1
10	2.2
6	6.5
8	7
10	4
16	2.5
18	1.6

Education (Yrs)	Jail Time (Yrs)
10	5.2
28	0.1
5	8.7
8	8.9
9	7.6
12	2.3
14	4.5
15	2.1
17	1.3
21	0.4
23	0.9
7	9.1

a. Determine the explanatory and response variables Education - expl.

b. Sketch a scatterplot of the data. Describe the scatterplot. Jail - resp.



c. Find the equation of the LSR line and the correlation coefficient. Sketch the LSR line on your scatterplot from (a).

$y = -0.412x + 9.599$ $\quad r = -0.875$

d. Use the model to predict the number of years in jail for someone with 18 years of education.

$y = -0.412(18) + 9.599$

$y = 2.177$ years

e. Calculate the residual for the prediction in part (d)

$1.6 - 2.177 = -0.577$

f. Is this prediction an overestimate or an underestimate?

g. Interpret the slope of the LSR line in a complete sentence.

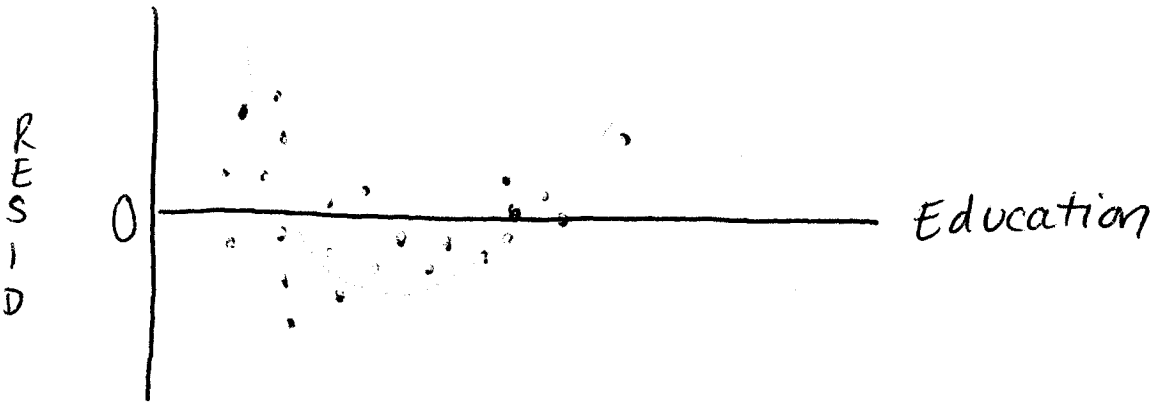
$\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{-0.412 \text{ years jail}}{1 \text{ years ed.}}$ For every 1 year increase in education, the # years in jail decreases by −0.412 years

h. Given that a person has spent 5 years in jail, how many years of education would you predict they have had?

$5 = -0.412x + 9.599$

$x = 11.16$ years

i. Sketch the residual plot.



j. What does the residual plot in part (i) tell us about our linear model? Justify.

not the best model, because residual plot is scattered.

k. Find the coefficient of determination and interpret it.

$r^2 = 0.766$

76.6% of the change in jail time is due to the change in education level.

l. What percent of jail time is due to factors OTHER than years of education?

100% - 76.6% = 23.4%

m. List some of these other factors that affect jail time (other than years of education). In other words, list some confounding/lurking variables in this situation.

- background
- where you live
- crime committed
- influences growing up
- wealth
- etc...

MULTIPLE CHOICE:

The stock market did well during the 1990s. Here are the percent total returns (change in price plus dividends paid) for the Standard & Poor's 500 stock index:

Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Return	31.7	−3.1	30.5	7.6	10.1	1.3	37.6	23.0	33.4	28.6

The next three questions are related to this situation.

1. The correlation of U.S. stock returns with overseas stock returns during these years was r = 0.44. This tells you that

(a) when U.S. stocks rose, overseas stocks also tended to rise, but the connection was not very strong
(b) when U.S. stocks rose, overseas stocks rose by almost exactly the same amount
(c) when U.S. stocks rose, overseas stocks tended to fall, but the connection was not very strong
(d) there is almost no relationship between changes in U.S. stocks and changes in overseas stocks
(e) nothing, because this is not a possible value of r

2. If x is the return on U.S. stocks and y is the return on overseas stocks in the same year, the least-squares regression line for predicting y from x is $y = -2.7 + 0.47x$. You think U.S. stocks will have a return of 10% in 1999. Using this regression line, you predict that the return on overseas stocks will be

(a) 7.4% (b) -2.23% (c) 2% (d) 3.17%

3. Stock returns are measured in percent. What are the units of the mean, the median, the quartiles, the standard deviation, and the correlation between U.S. and overseas returns?

(a) all are measured in percent.
(b) all are measured in percent except the standard deviation, which is measured in squared percent.
(c) all are measured in percent except the correlation, which is a number that has no units.
(d) all are measured in percent except the correlation, which is measured in squared percent.

5. Suppose that the correlation between the scores of students on Exam 1 and Exam 2 in a statistics class is r = 0.7. One way to interpret r is to say what percent of the change in Exam 2 scores can be explained by the change in Exam 1 scores. This percent is about

(a) 84% (b) 70% (c) 49% (d) 30%

7. What can we say about the relationship between a correlation r and the slope b of the least-squares line for the same set of data?
(a) r is always larger than b
(b) r and b always have the same sign (+ or -)
(c) b is always larger than r
(d) b and r are measured in the same units

13. Which statistical measure is **not** strongly affected by a few outliers in the data?
(a) the mean
(b) the median
(c) the standard deviation
(d) the correlation coefficient

16. The least-squares regression line for predicting the percent of a country's females who are illiterate from the percent of males who are illiterate is

female % = 3.34 + 1.39 × male %

In China, 10.1% of men are illiterate. Predict the percent of illiterate women in China.
(a) 4.7% (b) 14% (c) 17.4% (d) 47.8%

17. The equation of the regression line tells us that (on the average) when the male illiteracy rate goes up by 1%, the female rate goes up by
(a) 4.73% (b) 3.34% (c) 1.95% (d) 1.39%

19. You are planning an experiment to study the effect of gasoline brand and vehicle weight on the gas mileage (miles per gallon) of sport utility vehicles. In this study,
(a) gas mileage is a response variable.
(b) gas mileage is an explanatory variable.
(c) gas mileage is a lurking variable.
(d) gas mileage is a categorical variable.

21. A study of 3,617 adults found that those who attend religious services live longer (on the average) than those who don't. Is this good evidence that attending services *causes* longer life?
(a) Yes, because the study is an experiment.
(b) No, because religious people may differ from non-religious people in other ways, such as smoking and drinking, that affect life span.
(c) Yes, because the sample is so large that the margin of error will be quite small.
(d) No, because we can't generalize from 3,617 people to the millions of adults in the country.

22. Which of these is *not* true of the correlation r between the lengths in inches and weights in pounds of a sample of brook trout?
(a) r must take a value between -1 and 1.
(b) r is measured in inches.
(c) if longer trout tend to also be heavier, than $r > 0$.
(d) r would not change if we measured these trout in centimeters instead of inches.
(e) Both (b) and (d).

23. A correlation cannot have the value
(a) 0.4 (b) -0.75 (c) 1.5 (d) 0.0 (e) 0.99

24. Which correlation indicates a strong positive straight line relationship?
(a) 0.4 (b) -0.75 (c) 1.5 (d) 0.0 (e) 0.99

25. A study found that SAT verbal scores were positively associated with first-year grade point averages for liberal arts majors. We can conclude from this that
(a) students who scored high on the SAT verbal test tended to get lower GPAs than those who scored lower on the SAT verbal test
(b) students who scored high on the SAT verbal test tended to get higher GPAs than those who scored lower on the SAT verbal test
(c) we can use the SAT verbal score to accurately predict GPAs for liberal arts majors
(d) grade point averages are higher for older students
(e) the correlation between the SAT verbal score and GPA is higher than 0.5

30. If the least squares regression line for predicting y from x is $y = 500 - 20x$, what is the predicted value of y when $x = 10$?
(a) 300 (b) 500 (c) 200 (d) 700 (e) 20

31. Suppose that the least squares regression line for predicting y from x is $y = 100 + 1.3x$. Which of the following is a possible value for the correlation between y and x?
(a) 1.3 (b) -1.3 (c) 0 (d) -0.5 (e) 0.5

28. The correlation between two variables is of -0.8. We can conclude
(a) one causes the other
(b) there is a strong positive association between the two variables
(c) there is a strong negative association between the two variables
(d) all of the relationship between the two variables can be explained by a straight line
(e) there are no outliers

38. Perfect correlation means all of the following **except**
(a) $r = -1$ or $r = +1$.
(b) all points on the scatterplot lie on a straight line.
(c) all variation in one variable is explained by variation in the other variable.
(d) there is a causal relationship between the variables.
(e) each variable is a perfect predictor of the other.